Controllable thousand-port low-latency optical packet switch architecture for short link applications

S. Di Lucente[1], Y. Nazarathy[1,2], O. Raz[1], N. Calabretta[1] and H. J. S. Dorren[1]

[1] Eindhoven University of Technology, dept. Electrical Engineering, 5600MB Eindhoven, The Netherlands
[2] EURANDOM, Eindhoven University of Technology

The implementation of a low-latency optical packet switch architecture that is controllable while scaling to over thousand ports is investigated in this paper. Optical packet switches with thousand of input/output ports are promising devices to improve the performance of short link applications in Data Centers (DCs) or High Performance Computers (HPCs) environments. We show that classical switch architectures with a centralized control would not meet latency requirements of such applications. We present a WDM optical packet switch modular architecture with highly distributed control that allows scaling to over thousand ports while providing sub-microsecond end-to-end latencies.

Introduction

Current DCs consist of tens of thousands of interconnected servers [1], while in recent HPCs more than one hundred thousand processor cores are connected each other [2]. The extremely high number of needed interconnections and the fast operations that have to be provided within these environments lead to the need of interconnects that can scale to a large number of ports, while providing very low end-to-end latencies (≤1μs). Due to the limited port-count of actual interconnects, the servers in DCs and the processors in HPCs are interconnected according to a fat-tree topology. In this scenario, Optical Packet Switches (OPSs) that scale to thousands of ports would be of clear benefit, proving higher interconnection levels and allowing to interconnect servers (or processors) according to a flattened network topology. OPSs can provide higher bandwidth and lower end-to-end latencies, but scaling these devices to thousands of ports is still an open challenge. Fast OPSs that can handle high bit-rate input/output ports have already been presented [3-6], but despite of the decennial research studies in this field, OPSs with thousands of input/output ports do not exist. In this work, we focus on a promising WDM OPS architecture that features an highly distributed electronic control. As we will show, architectures that employ a centralized control would not meet the latency requirements while scaling to thousand ports. We investigate the performance of the proposed architecture in terms of latency, packet loss and throughput by means of network simulations. We show that end-to-end latencies below 1μs can be obtained in a flow controlled computer communication network, while scaling the switch architecture to over thousand ports. We also show that the studied architecture can provide relatively high throughput and low packet losses under high-load operations.

System under investigation and control issue

Figure 1 (a) shows a schematic of the system investigated in this paper. The switch architecture connects N inputs to N outputs. The nodes at the input, that can be regarded as a group of processors in a HPC or a group of servers of a DC, may contain queues in

electronic buffers. The switch is assumed to operate transparently in the optical domain, no opto-electrical conversions are needed in this device. However, the switch is electronically driven by the control. We supposed that the system operates in a synchronous mode, thus it operates in discrete time and all the requests of routing have to be handled at the same time by the switch control.
Consider the switch of Figure 1 (a) based on a Beneš architecture. The Beneš architecture is a rearrangeable non-blocking architecture in which every input can be connected to every output. However, at every time slot the entire switch matrix has to be reconfigured in order to connect the inputs with the right outputs. This process plays an important role when scaling the switch port-count to over thousand ports. The time to configure the connection map in a Beneš switch is port-count dependent and despite of multi-processor implementations, it scales at least linearly with the number of input/output ports of the switch [7-9]. This means that if the switch has 1000 ports it will take 1000 clock-cycles only to compute the new connection map every time slot in which the input state of the switch changes. Figure 1 (b) shows a switch with N ports based on a Spanke architecture.

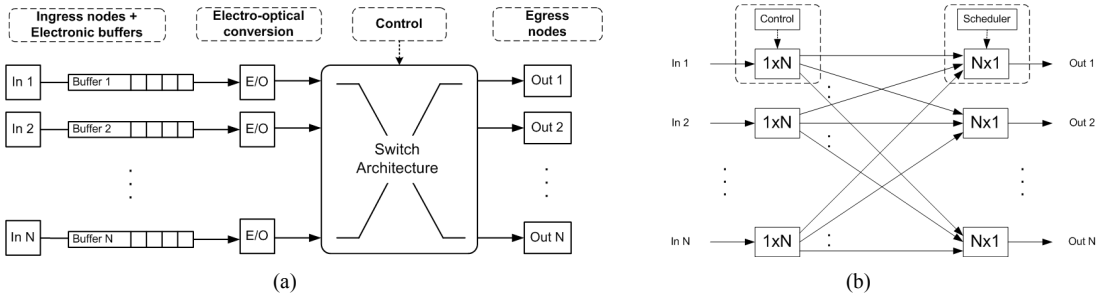


Figure 1: schematic of the system under investigation (a) and NxN switch based on Spanke architecture (b).

The switch of Figure 1 (b) is based on a strictly non-blocking architecture made of N 1xN and N Nx1 switches. These basic components can be autonomously controlled reducing the configuration time of the entire switching matrix to the time it takes to configure a single 1xN and a single Nx1 switch. For this reason the focus of our research is on Spanke-type switch architectures.

Proposed architecture: WDM and contention resolution

The Spanke architecture of Figure 1 (b) could provide very little latency. 1xN switches with 16 and 100 ports that can be controlled in few nanoseconds due to advanced labelling techniques and parallel label processors have already been presented [10, 11] . However, scaling these devices to over thousand ports is not feasible due to OSNR limitations [12]. The use of WDM technique in such devices will allow to scale to a large number of ports. The proposed architecture is depicted in Figure 2.
In the switch of Figure 2 there are N input fibers each of them carrying M wavelength channels. The NM logical inputs of the switch are fed into NM autonomously controlled 1xN switches. The label processors process the labels and drive the switches in order to forward the optical packets to the right output in few nanoseconds. Implementations of the label processor have already discussed in [13]. Obviously, packets on the same wavelength channel coming from different input fibers may contend the same output port. To solve these contentions, a modular contention resolution block has been implemented at each of the outputs of the switch. The M signals coming from the same

input fiber are grouped by a Wavelength Selector (WS) and connected to a Fixed Wavelength Converter (FWC). This means that only one packet of the same input fiber can be addressed to a single output port per time slot. The scheduler drives the WS and decides which packet has to be delivered. The FWCs convert the packets coming from different input fibers using different wavelengths in order to avoid contentions. For drawing simplicity, Figure 2 refers to a switch in which N=M, but there are not limitations in the number of wavelength channels that can be used in the system.

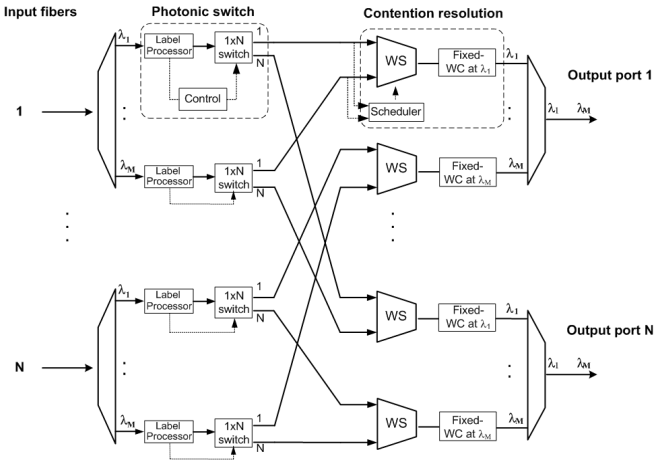
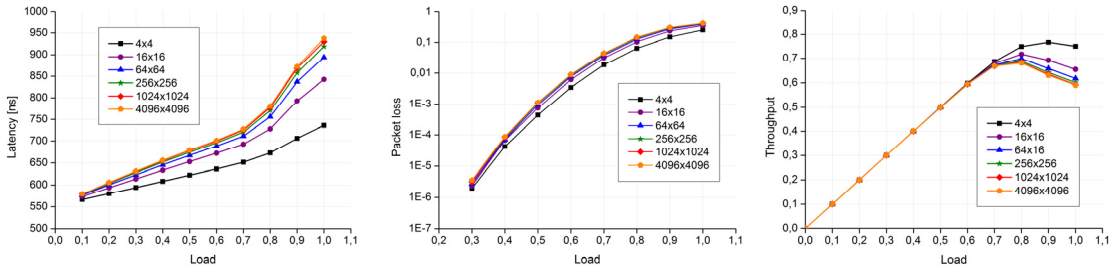


Figure 2: proposed architecture with contention resolution based on wavelength conversion at its output ports.

Simulation in a computer communication network

We investigate the performance in terms of latency, packet loss and throughput of the proposed architecture in a computer communication network with flow control by means of OMNeT++ simulations [14]. With flow control we mean that a packet can be considered successfully delivered only when an acknowledgement message is received back by the input. We consider a length link of 50 meters and a packet duration of 40 ns (considering 6 ns of guard band) and a bit rate of 40 Gb/s on each channel. We assumed a Bernoulli arrival pattern at each input with fixed probability (hereafter called load) and uniform destinations. Figure 3 (a), (b) and (c) show respectively latency, packet loss and throughput of the system of Figure 1 (a) implemented using the WDM OPS of Figure 2 and considering an input buffer size of 14 packets.



(a) (b) (c)

Figure 3: Latency (a), packet loss (b) and throughput (c) of the simulated system in function of load for switches based on the proposed architecture with 4, 16, 64, 256, 1024 and 4096 ports.

Figure 3 (a) shows that the latency increases with the load and it is port-count independent for switches with more than 256 input/output ports. It shows also that sub-microsecond end-to-end latencies can be achieved regardless the load and the port-

count. Figure 3 (b) shows that also the packet loss increases with the load and it is almost independent of the switch size. It is visible that a buffer size of only 14 packets can guarantee no losses for loads lower than 0.3. Finally, Figure 3 (c) shows that the throughput of the system increases linearly with the load and is independent of the port-count until loads around 0.7, then it starts decreasing faster for larger switch sizes. This behavior is due to the buffer overflowing. Under high loads operation, many packets have to be retransmitted and this phenomenon limits the throughput of the switch.

Conclusion

This work shows that OPSs with thousands of ports with a centralized control would not meet the latency requirements of applications in DC and HPC environments. We propose a promising WDM Spanke-type architecture with contention resolution based on wavelength conversion. We show by means of simulations that the proposed architecture can provide sub-microsecond end-to-end latency in a computer communication network regardless the input load and the port-count. We also show that low packet loss and high throughput can be achieved under relatively high load operations.

References

[1] L. A. Barros and U. Hőlzle, "The Datacenter as a Computer: An Introduction to the Design of Warehouse-scale Machines", synthesis lectures on Computer Architecture 4(1): pp. 1-118, 2009.

[2] A. KomornickiG et al. "Roadrunner: Hardware and software overview", IBM Redpaper, 2009.

[3] R. Luijten, W. E. Denzel, R. R. Grzybowski and R. Hemenway, "Optical interconnection network: The OSMOSIS project", Lasers and Electro-Optics Society, vol. 2, pp. 563-564, 2004.

[4] S.C. Nicholes, M.L. Mašanović, B. Jevremović, E. Lively, L.A. Coldren and D.J. Blumenthal, "The World's First InP 8x8 Monolithic Tunable Optical Router (MOTOR) Operating at 40 Gbps Line Rate per Port", Proceedings OFC 2009 post deadline paper B1, San Diego, USA, 2009.

[5] J. Gripp, M. Duelk, J. E. Simsarian, A. Bhardwaj, P. Bernasconi, O. Laznicka, and M. Zirngibl, "Optical switch fabrics for ultra-highcapacity IP routers", Journal of Lightwave Technology, vol. 21, pp. 2839-2850, 2003.

[6] H. Wang, A. Wonfor, K. A. Williams, R. V. Penty, I. H. White, "Demonstration of a Lossless Monolithic 16x16 QW SOA Switch", Proceedings ECOC 2009, post-deadline papers page 28 (PD 1.7), Vienna, Austria, 2009.

[7] D. C. Opferman and N.T. Tsao-Wu, "On a class of rearrangeable switching networks", Bell System Technical Journal, vol. 50(5), pp. 1579-1618, 1971.

[8] S. Andresen, "The looping algorithm extended to Base 2^t rearrange able switching networks", IEEE Transactions on Communications, COM-25(10), pp. 1057-1063, 1977.

[9] T. T. Lee, and S. Y. Liew, Parallel Routing Algorithms in Beneš-Clos Networks, IEEE Transactions on Communications, vol. 50(11), pp. 1841-1847, 2002.

[10] I. M. Soganci, T. Tanemura, K. A. Williams, N. Calabretta, T. de Vries, E. Smalbrugge, M. K. Smit, H. J. S. Dorren, and Y. Nakano, "Integrated Phased-Array 1x16 Photonic Switch for WDM Optical Packet Switching Application", Proceedings of the International Conference on Photonics in Switching 2009, pp. (WeI3-1/2), Pisa, Italy, 2009.

[11] I. M. Soganci, T. Tanemura, K. Takeda, M. Zaitsu, M. Takenaka and Y. Nakano, "Monolithic InP 100-port photonic switch", in Proc. Optical Communication (ECOC) 2010, Turin, Italy, 2010.

[12] R. Spanke, "Architecture for large nonblocking optical space switches". Quantum Electronics, IEEE Journal of 22(6), pp. 964-967, 1986.

[13] N. Calabretta, I. M. Soganci, T. Tanemura, W. Wang, O. Raz, K. Higuchi, K. A. Williams, T. J. de Vries, Y. Nakano, and H. J. S. Dorren, "1×16 optical packet switch sub-system with a monolithically integrated InP optical switch", in Proc. Optical Fiber Communications 2010, San Diego, USA, 2010.

[14] http://www.omnetpp.org/